The Open University

M248 Solutions to the Specimen Examination Paper

References to the *Handbook* and to units are given (where appropriate). A reference such as '*Handbook*, B1:6', for example, relates to point 6 in the unit outline for *Unit B1*.

PART 1

In many cases below, the working that leads to the correct option(s) is included. Please note that you will not need to provide any such working to answer Part 1 questions in the real examination.

Question 1 (*Unit A1*, Subsection 4.2)

Correct options: C and F (1 mark each).

The lower and upper quartiles are (*Handbook*, A1:5)

$$\begin{aligned} q_L = x_{\left(\frac{1}{4}(8+1)\right)} = x_{\left(2\frac{1}{4}\right)} &= x_{(2)} + \tfrac{1}{4}(x_{(3)} - x_{(2)}) \\ &= 1.07 + \tfrac{1}{4}(1.11 - 1.07) = 1.08 \end{aligned}$$

and

$$\begin{aligned} q_U = x_{\left(\frac{3}{4}(8+1)\right)} = x_{\left(6\frac{3}{4}\right)} &= x_{(6)} + \tfrac{3}{4}(x_{(7)} - x_{(6)}) \\ &= 1.14 + \tfrac{3}{4}(1.26 - 1.14) = 1.23. \end{aligned}$$

Question 2 (*Unit A2*, Subsection 1.1)

Correct option: B (2 marks).

For these data,

$$q_U + (1.5 \times iqr) = 39.5 + (1.5 \times 13.5) = 59.75.$$

The highest observation not exceeding 59.75 is 56.

So the upper adjacent value is 56.

Question 3 (*Unit A1*, Subsections 4.1–4.3)

Correct options: A, F and H (1 mark each).

 WEB 03989 3
4.1

Question 4 (*Unit D3*, Subsection 3.1)

Correct option: E (1 mark).

Proportion of accidents attributable to unsafe working conditions

$$= \frac{\text{Total number of 'Yes'}}{\text{Total number of accidents}} = \frac{46}{200} = 0.23.$$

Question 5 (*Unit D3*, Subsection 3.1)

Correct option: A (1 mark).

Proportion of accidents that occurred on the night shift

$$= \frac{\text{Total number of 'Night'}}{\text{Total number of accidents}} = \frac{54}{200} = 0.27.$$

Question 6 (*Unit D3*, Subsection 3.1)

Correct option: C (1 mark).

P(accident on night shift was due to unsafe working conditions)

$$= \frac{\text{Number of 'Yes' on the night shift}}{\text{Total number on the night shift}} = \frac{10}{54} = 0.185.$$

Question 7 (*Unit D3*, Subsection 3.1)

Correct option: F (1 mark).

P(accident that was attributable to unsafe working conditions occurred on the day shift)

$$= \frac{\text{Number of 'Yes' on the day shift}}{\text{Total number of 'Yes'}} = \frac{20}{46} = 0.435.$$

Question 8 (*Unit C3*, Subsection 1.2)

Correct option: B (1 mark).

There are four wheel bearings, which can be either damaged or not damaged, and the number of these four that are damaged, X, is of interest. As there are two possible outcomes for each wheel bearing, being damaged can be considered as a 'success' so that a binomial distribution (with $n = 4$) would be suitable for X.

Question 9 (*Unit C3*, Subsection 1.3)

Correct option: F (1 mark).

Y is a continuous variable. It is likely that the lengths of antennae would be clustered around a central value (which would suggest a normal distribution rather than a uniform distribution), and that this central value would be greater than 0 (which would suggest a normal distribution rather than an exponential distribution, which has a mode at 0). Therefore a normal distribution would be suitable for Y.

Question 10 (*Unit C3*, Subsection 1.2)

Correct option: D (1 mark).

Traffic accidents are occurring at random, and Z is the number of traffic accidents occurring during a fixed interval of time, namely each day, so a Poisson distribution would be suitable for Z.

Question 11 (*Unit C3*, Subsection 1.3)

Correct option: E (1 mark).

U is a (continuous) waiting time. It seems reasonable to assume that the arrivals arise at random, so an exponential distribution would be suitable for U.

Question 12 (*Unit A4*, Subsection 2.1)

Correct option: D (2 marks)

The mean is given by (*Handbook*, A4:3)

$$\begin{aligned} E(Y) &= \left(0 \times \tfrac{1}{8}\right) + \left(1 \times \tfrac{1}{4}\right) + \left(2 \times \tfrac{3}{8}\right) + \left(3 \times \tfrac{1}{4}\right) \\ &= \tfrac{1}{4} + \tfrac{3}{4} + \tfrac{3}{4} = 1\tfrac{3}{4}. \end{aligned}$$

Question 13 (*Unit A4*, Subsection 3.1; *Unit B1*, Subsection 3.1)

Correct option: B (2 marks).

The variance is given by (*Handbook*, A4:3)

$$\begin{aligned} V(X) &= \sum_x (x-\mu)^2\, p(x) \\ &= \sum_x \left(x - \tfrac{1}{6}\right)^2 p(x) \\ &= \left(-\tfrac{7}{6}\right)^2 \times \tfrac{1}{3} + \left(-\tfrac{1}{6}\right)^2 \times \tfrac{1}{6} + \left(\tfrac{5}{6}\right)^2 \times \tfrac{1}{2} = \tfrac{29}{36}. \end{aligned}$$

Alternatively (*Handbook*, B1:8),

$$\begin{aligned} V(X) &= E(X^2) - \mu^2 \\ &= E(X^2) - \left(\tfrac{1}{6}\right)^2 \\ &= (-1)^2 \times \tfrac{1}{3} + 0^2 \times \tfrac{1}{6} + 1^2 \times \tfrac{1}{2} - \tfrac{1}{36} = \tfrac{29}{36}. \end{aligned}$$

Question 14 (*Unit B2*, Subsection 3.2)

Correct options: D and E (1 mark each).

The number of defective fuses in a batch of 1000 has the binomial $B(1000, 0.02)$ distribution. So by the central limit theorem (*Handbook*, B2:6), the normal distribution that can be used to calculate the approximate value for the probability of observing 27 or more defective fuses in a random sample of 1000 fuses has mean and variance given by

$$\text{mean} = 1000 \times 0.02 = 20,$$

$$\text{variance} = 1000 \times 0.02 \times 0.98 = 19.6.$$

Question 15 (*Unit A5*, Section 1)

Correct option: A (2 marks).

Let X be the number of calls arriving per hour. Then $X \sim \text{Poisson}(4)$, so (*Handbook*, A5:1)

$$P(X = 2) = \frac{e^{-4}4^2}{2!} \simeq 0.147.$$

Question 16 (*Unit A5*, Section 1)

Correct option: E (2 marks).

For $X \sim \text{Poisson}(4)$, the required probability is (*Handbook*, A5:1)

$$\begin{aligned} P(X \geq 2) &= 1 - P(X = 0) - P(X = 1) \\ &= 1 - \frac{e^{-4}4^0}{0!} - \frac{e^{-4}4^1}{1!} \simeq 0.908. \end{aligned}$$

Question 17 (*Unit A5*, Section 3)

Correct option: C (2 marks).

The interval in hours between successive calls, T say, has an exponential distribution with parameter $\lambda = 4$ (*Handbook*, A5:6). Ten minutes is $\frac{1}{6}$ of 60 minutes, so

$$P(T < 10 \text{ minutes}) = P\left(T < \tfrac{1}{6}\right) = 1 - e^{\left(-4 \times \frac{1}{6}\right)} \simeq 0.487.$$

Question 18 (*Unit A5*, Section 3)

Correct option: F (2 marks).

There are two hours between 9.00 am and 11.00 am. So the number of calls arriving in two hours has a Poisson distribution with parameter (*Handbook*, A5:6)

$$(4 \text{ per hour}) \times 2 = 8.$$

Question 19 (*Unit A3*, Section 5)

Correct option: F (2 marks).

The probability required is (*Handbook*, A3:12)

$$P(Z < 5) = P(Z \leq 4) = F(4) = 1 - q^4 = 1 - (0.3)^4 = 0.9919.$$

Question 20 (*Unit B2*, Subsection 2.4)

Correct option: A (2 marks).

The proportion of 100-tile packs that weigh more than 10 kilograms is given by (*Handbook*, B1:5)

$$P(Y > 10) \simeq P\left(Z > \frac{10 - 9.9}{0.05}\right) = 1 - \Phi(2) = 1 - 0.9772 = 0.0228.$$

Question 21 (*Unit B2*, Subsection 2.4)

Correct option: A (2 marks).

Let T_{100} be the total weight of 100 passengers. Then T_{100} is approximately normal with (*Handbook*, B2:4)

$$E(T_{100}) = 100 \times 70 = 7000 \quad \text{and} \quad V(T_{100}) = 100 \times 64 = 6400.$$

Then

$$\begin{aligned} P(T_{100} < 7200) &= P\left(Z < \frac{7200 - 7000}{\sqrt{6400}}\right) \\ &= P(Z < 2.5) = 0.9938 \simeq 0.994. \end{aligned}$$

Question 22 (*Unit A5*, Subsection 2.2)

Correct option: B (2 marks).

$X \sim M(2)$, so (*Handbook*, 4: Table of continuous probability distributions)

$$V(X) = \frac{1}{2^2} = \frac{1}{4}.$$

Thus

$$V(W) = V(4X) = 16 \times V(X) = \frac{16}{4} = 4.$$

Question 23 (*Unit C2*, Subsection 2.2)

Correct option: B (2 marks).

$X \sim \chi^2(5)$, so (*Handbook*, 4: Table of continuous probability distributions)

$$V(X) = 2 \times 5 = 10.$$

Also, $Y \sim \chi^2(3)$, so

$$V(Y) = 2 \times 3 = 6.$$

Then (*Handbook*, B1:11)

$$SD(U) = \sqrt{V(X - Y)} = \sqrt{V(X) + V(Y)} = \sqrt{10 + 6} = \sqrt{16} = 4.$$

Question 24 (*Unit B3*, Subsection 3.1)

Correct option: B (2 marks).

If μ denotes the mean birth weight in pounds, and θ denotes the mean birth weight in grams, then $\theta = 454 \times \mu$.

This is an increasing function, so the corresponding 95% confidence interval for θ is (*Handbook*, B3:3)

$$(454 \times 6.87, 454 \times 7.45) = (3118.98, 3382.30).$$

Question 25 (*Unit B3*, Subsection 3.2)

Correct option: E (2 marks).

An estimate of p is $\widehat{p} = 6/3000 = 0.002$. The upper limit of an approximate 95% confidence interval for p is (*Handbook*, B3:4)

$$\widehat{p} + z \times \sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}} = 0.002 + 1.96 \times \sqrt{\frac{0.002 \times 0.998}{3000}} = 0.0036.$$

Question 26 (*Unit B3*, Subsection 5.1)

Correct option: C (2 marks).

A confidence interval for a normal mean is a t-interval (*Handbook*, B3:7).

For a 90% t-interval and sample of size 15, the 0.95-quantile of $t(n-1) = t(14)$ is needed; this is 1.761. The lower limit of the t-interval is then

$$\overline{x} - t \times \frac{s}{\sqrt{n}} = 12.5 - 1.761 \times \frac{4.4}{\sqrt{15}} \simeq 10.5.$$

Question 27 (*Unit C1*, Subsection 6.2)

Correct option: A (2 marks).

The required sample size is given by (*Handbook*, C1:12)

$$n = \frac{\sigma^2}{d^2}(q_{1-\alpha/2} - q_{1-\gamma})^2,$$

where $\sigma = 5$, $d = 2$, $q_{1-\alpha/2} = q_{0.975} = 1.960$ and $q_{1-\gamma} = q_{0.2} = -0.8416$. So

$$n = \frac{5^2}{2^2}(1.960 + 0.8416)^2 \simeq 49.06,$$

which is rounded up to 50.

Question 28 (*Unit D2*, Subsection 2.3)

Correct options: F and H (1 mark each).

S_{xx} and S_{xy} are calculated as follows (*Handbook*, D2:4):

$$S_{xx} = \sum x_i^2 - \frac{(\sum x_i)^2}{n} = 127 - \frac{26^2}{8} = 42.5,$$
$$S_{xy} = \sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n} = 598 - \frac{26 \times 180}{8} = 13.$$

Question 29 (*Unit D2*, Subsection 2.3)

Correct option: E (1 mark).

The least squares estimate of the slope parameter β for the regression line is (*Handbook*, D2:5)

$$\widehat{\beta} = \frac{S_{xy}}{S_{xx}} = \frac{284}{452} = 0.628 \simeq 0.63.$$

PART 2

*In Part 2 it is very important that you show **ALL** of your working.*

Question 30 (*Unit A2*, Section 1)

Two or more of the following comments will gain full marks.

Boxplots are much better for comparing several groups at the same time; boxplots display certain numerical summaries (median, quartiles), whereas histograms do not; boxplots give precise locations of outliers, whereas histograms display them only approximately; there are no cutpoint or interval width specifications for boxplots (and choosing these can be problematic for histograms). [*2*]

Question 31 (*Unit A1*, Subsection 3.1)

The cutpoints between groups are different in the two histograms. [*1*]

Question 32 (*Unit A3*, Subsection 3.1; *Handbook*, A3:5)

The function is not a probability mass function because $p(5)$ is negative ($p(5) = -\frac{1}{5}$). [*1*]

Question 33 (*Unit B3*, Subsections 2.1 and 2.2; *Handbook*, B3:1)

(a) If a large number of random samples of men <u>of size 529</u> were drawn independently from the population of men, and on each occasion a 90% confidence interval were calculated for the mean number of units consumed in the previous week, then approximately 90% of these intervals would contain the true mean number of units consumed by men in the previous week. [*1*]

(b) This confidence interval defines a plausible range for the mean number of units consumed in the previous week.

For example, if the true mean number of units consumed were greater than 20.7, then the probability of observing a sample mean less than or equal to 18.2 would be less than 0.05.

(Or, if the true mean number of units consumed were less than 15.7, then the probability of observing a sample mean greater than or equal to 18.2 would be less than 0.05.) [*2*]

Question 34 (*Unit C1*, Sections 1 and 2)

(a) The null hypothesis would be rejected at the 5% significance level.

These data provide evidence that the population mean is not equal to 3. [*2*]

(b) The p value in the first analysis must be less than 0.05. So either the first analysis must be incorrect, or the p value of 0.15 in the second analysis must be wrong (or both). [*1*]

Question 35 (*Unit A2*, Section 1; *Unit C2*, Subsection 1.3)

(a) The samples are right-skew. [*1*]

(b) The median is higher for the sample from population A, but the spread of values in the two samples is similar (as measured by the interquartile range). (You could have mentioned that the spread as measured by the range is greater for the sample from population B.) [*2*]

(c) (*Handbook*, C1:10) Since the samples are right-skew, it cannot be assumed that the populations are normally distributed. [*1*]

(d) (*Handbook*, C2:4)

(i) The null hypothesis is that the distributions of the populations from which the samples are drawn are the same. [*1*]

(ii) Under the null hypothesis,

$$\begin{aligned} U_A &\approx N\left(\frac{n_A(n_A+n_B+1)}{2}, \frac{n_A n_B(n_A+n_B+1)}{12}\right) \\ &= N(510, 2550). \end{aligned}$$

The p value is

$$\begin{aligned} 2\times P(U_A \geq 646) &\simeq 2\times P\left(Z \geq \frac{646-510}{\sqrt{2550}}\right) \\ &\simeq 2\times\left(1-\Phi(2.69)\right) \\ &= 0.0072. \end{aligned}$$

[*3*]

(iii) There is strong evidence that the populations from which the two samples are drawn are not the same. [*2*]

Question 36 (*Unit C1*, Subsection 4.2; *Handbook*, C1:10)

The hypotheses are

$$H_0\colon \mu_A = \mu_B, \quad H_1\colon \mu_A \neq \mu_B.$$

The null distribution of the test statistic is $t(n_1+n_2-2) = t(8)$.

The observed value of the test statistic is

$$t = \frac{\overline{x}_A - \overline{x}_B}{s_p\sqrt{\frac{1}{n_A}+\frac{1}{n_B}}} = \frac{1.92-3.12}{\sqrt{2.09\left(\frac{1}{5}+\frac{1}{5}\right)}} \simeq -1.312.$$

The 0.9-quantile of $t(8)$ is 1.397, so the p value must be greater than $2\times 0.1 = 0.2$. Hence there is insufficient evidence to reject the null hypothesis that the drugs produce the same mean effects. [*5*]

Question 37 (*Unit C3*, Section 4; *Handbook*, C3:3)

Statement	Section
1	C: Results
2	B: Methods
3	C: Results
4	B: Methods
5	A: Introduction
6	D: Discussion

[3]

Question 38 (*Unit D1*, Subsection 3.2; *Handbook*, D1:8)

(a) The MLE of θ is 8.29, the maximum value. *[1]*

(b) The maximum value in a sample is a biased estimator of θ (it underestimates θ). *[1]*

Question 39 (*Unit D1*, Subsection 3.1; *Handbook*, D1:4)

The likelihood of p for the sample is given by

$$L(p) = (1-p)^2 p \times p \times (1-p)^8 p = (1-p)^{10} p^3.$$ *[3]*

Question 40 (*Unit D2*, Subsection 4.1)

(a) A 1000 euro increase in net disposable income corresponds to 21.65 extra television sets per 1000 people (that is, $\widehat{\beta}$). *[1]*

(b) (*Handbook*, D2:8) The distribution of β is given by

$$\frac{\widehat{\beta}-\beta}{S/\sqrt{S_{xx}}} \sim t(n-2).$$

For a sample of size 11 and a 95% confidence interval, the 0.975-quantile of $t(9)$ is required; this is 2.262. So the confidence interval is

$$\begin{aligned}(\widehat{\beta} \pm 2.262 \times s/\sqrt{S_{xx}}) &= (21.65 \pm 2.262 \times \sqrt{1072.0/82.01}) \\ &\simeq (13.47, 29.83).\end{aligned}$$ *[4]*

Question 41 (*Unit D2*, Section 3)

Two plots should be obtained: a normal probability plot of the residuals, and a residual plot of residuals against fitted values.

The points on the normal probability plot should lie roughly along a straight line.

The points on the residual plot should appear to be randomly located, and the variance should not appear to vary with the fitted values. *[3]*

Question 42 (*Unit D3*, Section 2; *Handbook*, D3:3,4)

(a) Any scatterplot showing the points lying on a non-linear, monotonic increasing curve would obtain the mark. [*1*]

(b) If the Pearson correlation is 1, then the points lie on a straight line with positive slope, so the Spearman correlation must also be 1. [*1*]

Question 43 (*Unit D3*, Section 3)

(a) (*Handbook*, D3:8) The null hypothesis is that there is no association between experiencing jet lag and being given the treatment. [*1*]

(b) The expected frequencies for the cells are as follows.

	Jet lag	No jet lag
Treatment	6.5	8.5
Placebo	6.5	8.5

Hence the observed value of the test statistic is

$$\begin{aligned}\chi^2 &= \frac{(3-6.5)^2}{6.5} + \frac{(10-6.5)^2}{6.5} + \frac{(12-8.5)^2}{8.5} + \frac{(5-8.5)^2}{8.5} \\ &= \frac{12.25}{6.5} + \frac{12.25}{6.5} + \frac{12.25}{8.5} + \frac{12.25}{8.5} \\ &\simeq 6.65.\end{aligned}$$ [*2*]

(c) There is $(2-1)(2-1) = 1$ degree of freedom. The 0.99-quantile of $\chi^2(1)$ is 6.63, and the 0.995-quantile is 7.88. So the p value is between 0.005 and 0.01. [*2*]

(d) There is strong evidence against the null hypothesis that there is no association between experiencing jet lag and being given the treatment. The data suggest that the treatment suppresses jet lag. [*2*]